NISTIR 4747

Representations in Visual Motion

Daniel Raviv

Florida Atlantic University
The Robotics Center and
The Electrical Engineering Department
Boca Raton, FL 33431
and
Sensory Intelligence Group
Robot Systems Division

James S. Albus
Chief, Robot Systems Division

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Robot Systems Division
Bldg. 220 Rm. B124
Gaithersburg, MD 20899

U.S. DEPARTMENT OF COMMERCE
Rockwell A. Schnabel, Acting Secretary

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
John W. Lyons, Director

NIST

NISTIR 4747

Representations in Visual Motion

Daniel Raviv

Florida Atlantic University
The Robotics Center and
The Electrical Engineering Department
Boca Raton, FL 33431
and
Sensory Intelligence Group
Robot Systems Division

James S. Albus
Chief, Robot Systems Division

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Robot Systems Division
Bldg. 220 Rm. B124
Gaithersburg, MD 20899

January 1992

U.S. DEPARTMENT OF COMMERCE
Rockwell A. Schnabel, Acting Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
John W. Lyons, Director

ABSTRACT

The problem of constant perception despite varying visual sensations, i.e., how despite different images of the same scene, the world is perceived as stationary, has puzzled many researchers in the last five decades. In an attempt to answer this question another basic question arises: Is there an image-sequence transformation under which permanent properties of the environment are preserved during the motion of the eye?

This paper deals with motion-based image transformations that lead to new representations of 3-D objects for a rectilinear motion of an observer. In the new representations a stationary environment is kept invariant or "frozen" in spite of the fact that the images on the retina are continuosly changing. Since the 3-D stationary environment is also *represented* as a stationary environment, moving objects can be easily detected and a good estimation of the scene from a long sequence of noisy images can be obtained.

Three basic observations which are related to translational motion of an observer led to the new representations. One is that the radial distance of a point in 3-D space from the camera translational path is kept constant at any instant of time. The second observation is that the relative depth along the translational path between two points is the same at any point in time. The third observation is that points in the image plane move away from the Focus of Expansion (FOE) and towards the Focus of Contraction (FOC).

The new representations are simple. All measurements for these representations take place in camera coordinates, only one camera is necessary, and the magnitude of the camera velocity vector need not be known. The representations are pixel based, i.e., they involve local and parallel computations. The mathematics involved is simple, and thus it may be suitable for real time applications.

1. INTRODUCTION

During eye motion in a stationary environment, the projection of objects in the world is continuously changing (Figure 1), but we perceive the world as stationary. Why? Are there properties of the image that under some transformation remain constant during the motion of the eye? In other words, are there *visual invariants* that will result in new representations in which the transformed projected objects are "frozen"? Many researchers, especially in the field of psychology (e.g., Cutting [9] and Gibson [12]) have asked this question before and tried to answer it. If such invariants exist then there may be support for a new theory of perception.

Clearly, the geometrical relationships of a stationary environment are invariants. The problem is, why do we perceive them as such. What is measured and calculated such that it is perceived as stationary?

This paper deals with motion-based image transformations that lead to new representations of 3-D objects for a translational motion of an observer. In the new representations a stationary environment is kept invariant or "frozen" in spite of the fact that the images on the retina are continuously changing. The representations are based on nonlinear functions of optical flow, i.e., the coordinate system is optical flow dependent.

Three basic observations which are related to translational motion of an observer led to the new representations. One is that the radial distance of a point in 3-D space from the camera translational path is kept constant at any instant of time. The second observation is that the relative depth along the translational path between any two chosen points is the same at all instants of time. The third observation is that points in the image plane move away from the Focus of Expansion (FOE) and towards the Focus of Contraction (FOC).

The three coordinates for some of the new representations are: 1) the physical distance of a point in 3-D from the axis of motion expressed in terms of optical flow, 2) the physical depth of the point in 3-D from the camera pinhole point expressed in terms of optical flow, and 3) the angle of the radial line on which the point is moving in the

image plane.

The optical flow analysis takes place in a spherical $(R-\Theta-\Phi)$ coordinate system (rather than X-Y-Z). In this system, since we deal with translation only along the optical axis, any point in this coordinate system moves along a constant ϕ radial line and can be processed independently of any other point.

The new representations are simple. All the related measurements take place in camera coordinates, only one camera is used, and the magnitude of the camera velocity vector need not be known in some of the representations. The representations are pixel based, i.e., they involve local computations that can be executed in parallel.

We start with a description of the coordinate system (Section 2). In Section 3 it is shown how points in 3-D space share the same time dependent optical-flow based invariants. In Section 4 we show how to use the invariants to represent objects in 3-D space in a way such that they appear "frozen". Several domains are described.

2. COORDINATE SYSTEM FOR 3-D SPACE

In a rectilinear motion with no rotation, points in the image plane move radially away from the Focus of Expansion (FOE) (Figure 2a) and towards the Focus of Contraction (FOC) (Figure 2b). Based on this observation we use an R-Θ-Φ spherical (also known as the velocity egosphere [8]) rather than an X-Y-Z cartesian representation of points in space, which reduces to a Θ-Φ representation in the image domain.

Figures 3a, 3b, and 4 show the chosen coordinate system and the definitions of r and the angles θ and ϕ. If the optical axis coincides with the translational vector then, in the image domain constant ϕ corresponds to a radial line that emerges from the FOE and constant θ corresponds to a circle whose center is the FOE. Given a point $(x-y-z)$ in the $(X-Y-Z)$ coordinate system, it can be transformed to a $(r-\theta-\phi)$ point in the $(R-\Theta-\Phi)$ domain (i.e., the velocity egosphere [8]) and vice versa.

In this paper we assume that the camera (observer) undergoes only translational motion along the optical axis. Using the $R-\Theta-\Phi$ coordinate system, any point in the image domain moves radially, i.e., along a constant ϕ line. In the velocity egosphere coordinate system a point in the image can be processed independently of any other

point, and a constant ϕ line may be processed as a 1-D image.

3. MOTION INVARIANTS

We describe two invariants. (See [5] and Figure 5). It can be shown (see for example [4,8]) that

$$V = r\frac{\dot{\theta}}{\sin\theta} \tag{1}$$

where the dot denotes differentiation with respect to time. $\dot{\theta}$ is the optical flow along a radial line with constant ϕ (a great circle on the velocity egosphere). The value $r\frac{\dot{\theta}}{\sin\theta}$ is the same for *all points* (except for two singular points at $\theta=0$ and $\theta=\pi$) in 3-D space. If in addition the velocity V of the observer remains constant, then this relationship holds for all time instants. θ and $\dot{\theta}$ can be measured/computed at each instant of time, and hence the nonlinear function $\frac{\dot{\theta}}{\sin\theta}$ can be obtained.

The following is a description of the two invariants in terms of optical flow (see [5]). We denote them by $\frac{1}{\tau_C}$ and $\frac{1}{\tau_P}$. They all have units of $\frac{1}{time}$.

1. *The clearance invariant* $\frac{1}{\tau_C}=\frac{\dot{\theta}}{\sin^2\theta}$ (Figure 6): All points in 3-D space (except those that lie on the motion axis, i.e., $\theta=0$ and $\theta=\pi$) that lie on a cylindrical surface whose symmetry axis coincides with the camera motion direction share this invariant, i.e., have the same τ_C. In this case the radius of the cylinder $r\sin\theta$ is constant and so, using Equation (1), $\frac{\dot{\theta}}{\sin^2\theta}$ remains constant.

This invariant has been used by Raviv [3] to develop a robust, integration-based, and massively parallel method for reconstructing 3-D scenes. Albus [8] has suggested that this invariant can be used to measure clearance.

2. *The time to contact invariant* $\frac{1}{\tau_P}=\frac{2\dot{\theta}}{\sin 2\theta}$ (Figure 7): All points in 3-D space (except for $\theta=0$ and $\theta=\pi$) that lie on a plane which is perpendicular to the direction of motion of the camera, share this invariant, i.e., have the same τ_P. In this

case $r\cos\theta$ (the distance from the surface) is constant and so, using Equation (1), $\frac{2\dot{\theta}}{\sin 2\theta}$ remains constant.

This is the same invariant described by Lee [1] and his collegues [2]. More about extraction of τ_P from visual data can be found in [4].

4. THE NEW 3-D REPRESENTATIONS

We describe space-dependent and time-dependent representations. All representations are derived for an instantaneous translation of the observer in a stationary environment.

As mentioned in the introduction section of this paper there are three basic observations that led to the new representations. For a translational motion of of an observer the distance of a point in 3-D space from the camera translational path is constant at any instant of time, the relative depth between points is the same at any point in time, and points in the image plane move away from the Focus of Expansion (FOE).

SPACE REPRESENTATIONS

These representations are based on the two previously described invariants $\frac{1}{\tau_C}$ and $\frac{1}{\tau_P}$. Figure 8 shows one of the new coordinate systems. The axes are: $\tau_C V = \frac{\sin^2\theta}{\dot{\theta}}$, $\tau_P V = \frac{\sin 2\theta}{2\dot{\theta}}$ and Φ. $\tau_C V$ of a point is the distance from the point to the motion axis. $\tau_P V$ of a point is its depth (where the camera's focal point is at the origin). ϕ is constant.

Clearly, based on the previously described invariants and coordinate system a point in 3-D space which is represented in this space has the same $\tau_C V$, and the same ϕ for all time instants. Thus a 3-D surface appears the same in this coordinate system at all instants of time (except that it is shifted along the $\tau_P V$ axis). Note that in this representation any constant ϕ plane can be processed independently of other constant ϕ plane. The problem is reduced from 3-D to 2-D. Figure 9 shows a 3-D object in camera coordinates (depth, radial distance, and radial angle) at three different time instants

(figure 9a) and the representation of this object as extracted from the optical flow at the same three time instants (Figure 9b). Note that in the new domain the object appears the same at the three instants of time (but shifted vertically).

Another way of representing the information about a 3-D point is by having a coordinate system which is basically $X-Y-Z$ but expressed in terms of optical flow. The coordinates in this case are $\tau_C V\cos\phi$, $\tau_C V\sin\phi$, and $\tau_P V$. Figure 10a shows a 3-D object in the $X-Y-Z$ camera coordinate system at six different instants of time. Figure 10b shows the images of this object at the same six time instants. When using this optical flow based transformation the object appears as in Figure 10a, i.e., the same at all time instants but shifted along the vertical axis (Figure 10c).

TIME REPRESENTATION

The space representations depend on the knowledge of the magnitude of the vector V. However, in many cases, the speed V may not be available or not important (in behavior related situations the important parameters are the ratios between space and time for example the "time-to contact" [1]). In such cases a representation which is only time dependent can be used. The axes in this representations are τ_C, τ_P, and Φ or, $\tau_C\cos\phi$, $\tau_C\sin\phi$, and τ_P. The representation of objects in this case is scaled as a function of the velocity of the observer (Figure 11).

REPRESENTATIONS FOR "FREEZING" 3-D SURFACES

Based on the previously described space representations we describe other representations that freeze(!) the 3-D stationary environment while the camera is moving. In other words, any surface in space appears at the same location, orientation and size for all time instants. These representations are obtained by modifying the space representations. In the space representations a surface in 3-D space is invariant but is shifted along one axis as a function of time. The shift is actually the distance that the camera undergoes, or $-\int V(\tau)d\tau$ where the limits of the integral are from some initial time t_0 to the "current" time t_1.

The axes of these representations are (Figure 12): $\tau_C V$, $\tau_P V+\int V(\tau)d\tau$, and Φ, or $\tau_C V\cos\phi$, $\tau_C V\sin\phi$, and $\tau_P V+\int V(\tau)d\tau$. A 3-D object and its representation in the new

domain are shown.

Clearly, these representations are very useful for recursive estimation, Kalman filtering etc. They rely on local operations and involve simple computations. They can be used to eliminate global flow i.e., flow induced by the camera motion, detect local motion, avoid moving objects, robustly reconstruct 3-D objects with only one moving camera, recognize 3-D objects, etc.

5. RESULTS

Using a simulation program we verified one of the new representations. Figure 13 shows the projections of four identical 3-D objects at two different time instants. The projections were obtained at two different locations of the camera that underwent translation only. Note that the projections at the two time instants are completely different. The arrows show the translation of points on the objects along radial lines. At each of the two time instants we computed the coordinates of the eight points of each object in the new domain and verified that in the $\{\tau_C V, \tau_P V\}$ coordinates the computed values are obtained. Since $\dot{\theta}$ can be only approximated, the values that we obtained were "almost" the same.

6. FUTURE WORK AND CONCLUSIONS

Currently new representations are sought. We are trying to extend the representations to six-degree-of-freedom motion of the camera. The representations are being implemented on a real robotic system.

7. ACKNOWLEDGEMENT

The authors would like to thank Marty Herman, Ernie Kent and Don Orser for fruitful discussions regarding the invariants and the related representations.

8. REFERENCES

[1] D. N. Lee, "A theory of visual control of braking based on information about time to contact", Perception, No. 5, pp. 437-459.

[2] D. N. Lee and P. E. reddish, "Plumming gannets: a paradigm of ecological optics", Nature 293, pp. 293-294.

[3] D. Raviv, "Parallel algorithm for 3-D surface reconstruction", In proceedings of the SPIE conference on intelligent robots, Philadelphia, PA, Nov. 1989.

[4] D. Raviv, "Extracting the 'time to contact' parameters from real visual data", In proceedings of the SPIE conference on intelligent robots, Philadelphia, PA, Nov. 1989.

[5] D. Raviv, "Invariants in visual motion", NISTIR to be published.

[6] Raviv, D. and Herman, M., "Towards an understanding of camera fixation", IEEE Conference on Robotics and Automation, Cincinatti, Ohio, 1989.

[7] Raviv, D., "A quantitative Approach to Camera Fixation", In Proc. Computer Vision and Pattern Recognition (CVPR) Conference, Hawaii, June 1991.

[8] Albus, J.S., "Outline for a Theory of Intelligence", IEEE Transactions on systems, Man and Cybernetics, Vol. 21, No. 3, 1991.

[9] Cutting, J., *Perception with an Eye for Motion,* MIT Press, 1986.

[10] Raviv, D., "A Quantitative Approach to Looming and Fixation", NISTIR, in preparation.

[11] Raviv, D., and M. Herman, "A New approach to Vision and Control for Road Following", NISTIR 4476-91, January 91.

[12] Gibson, J.J., "New Reasons for Realism", Synthesis 17, pp. 162-172, 1967.

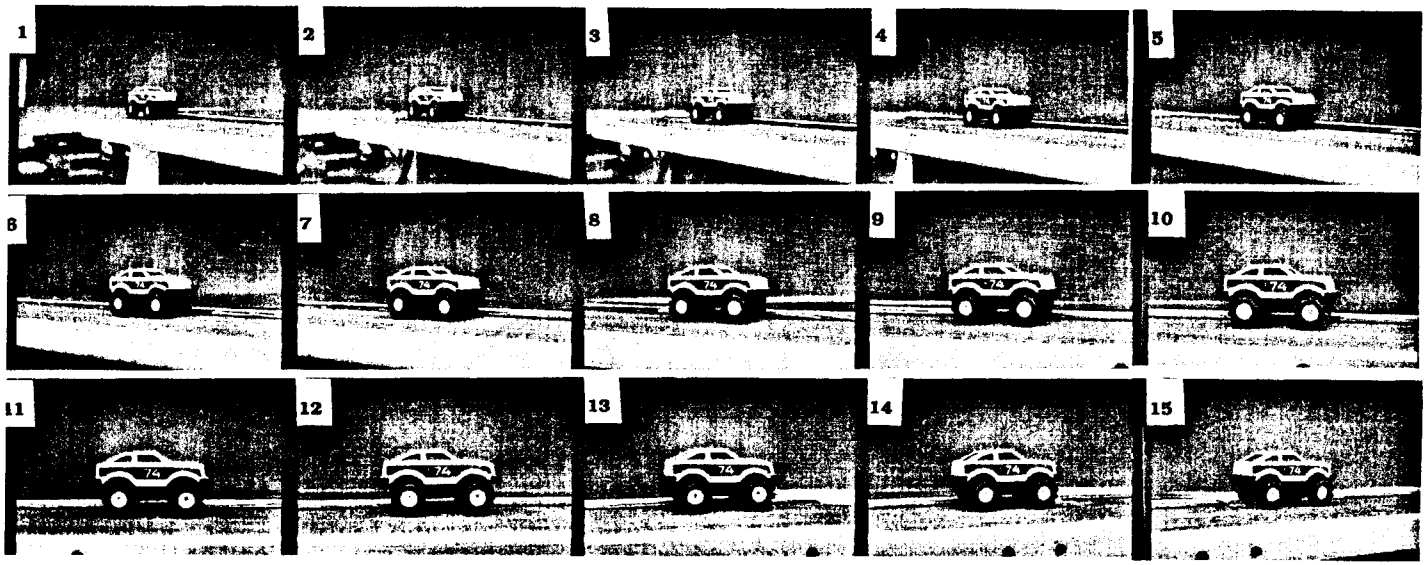


Figure 1: A Sequence of Images

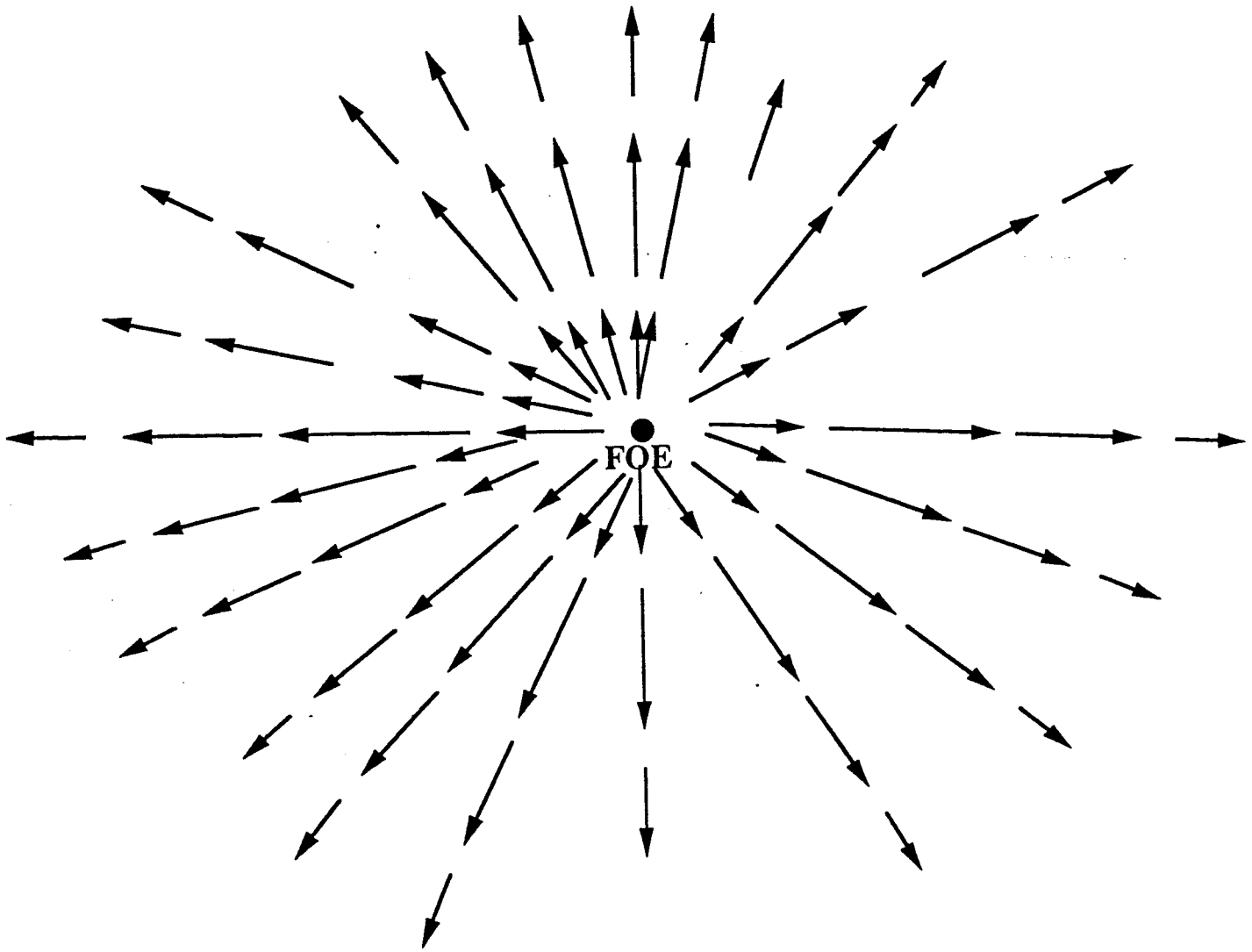


Figure 2a: Optical Flow Relative to the Focus of Expansion

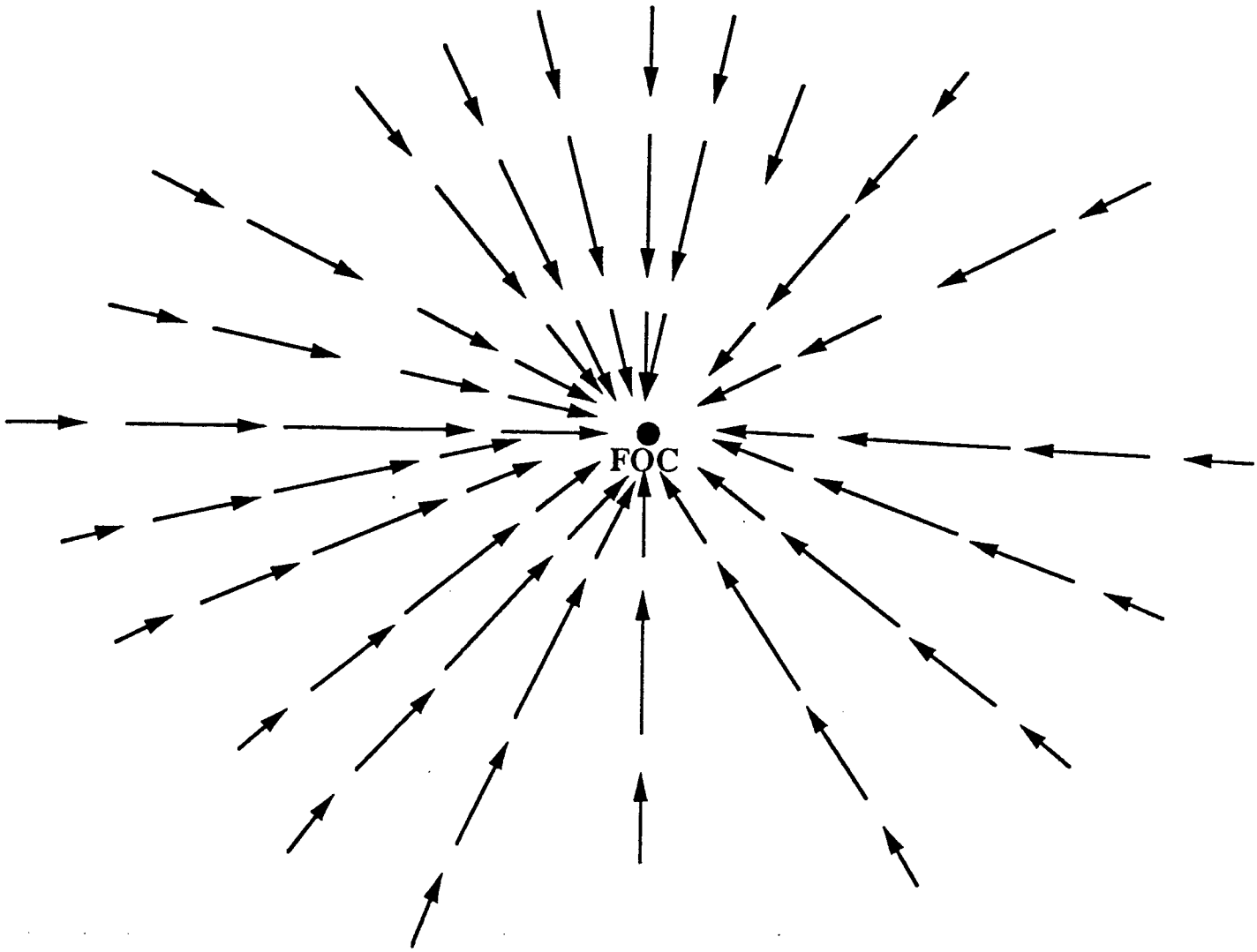


Figure 2b: Optical Flow Relative to the Focus of Contraction

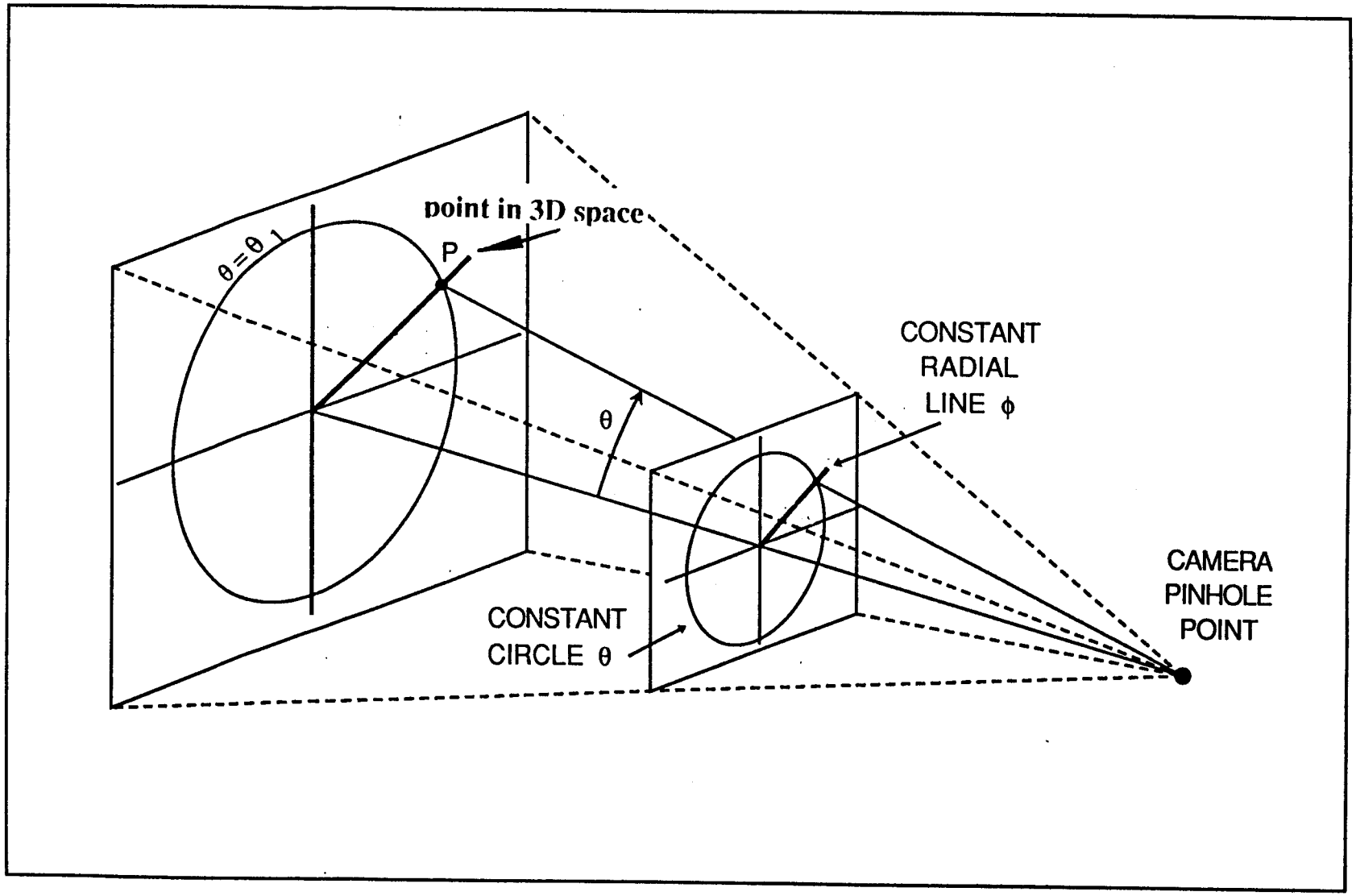


Figure 3a: Coordinate System

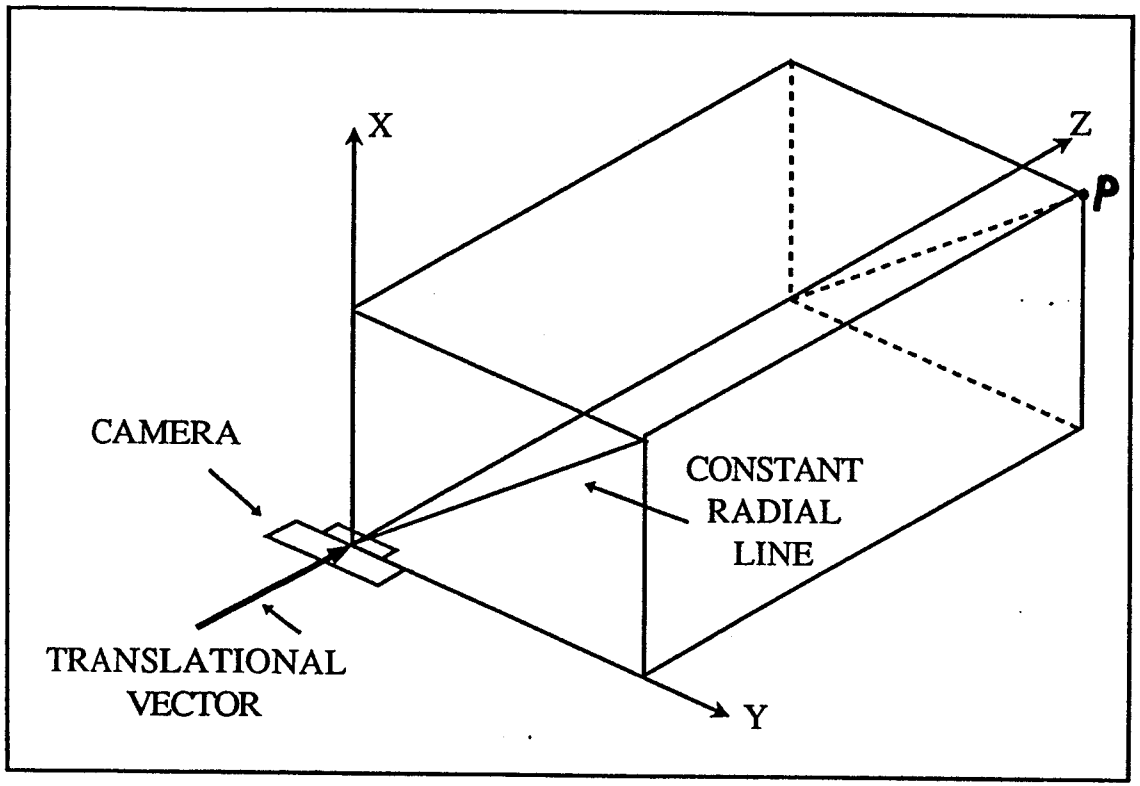


Figure 3b: Motion Along the Z-Axis

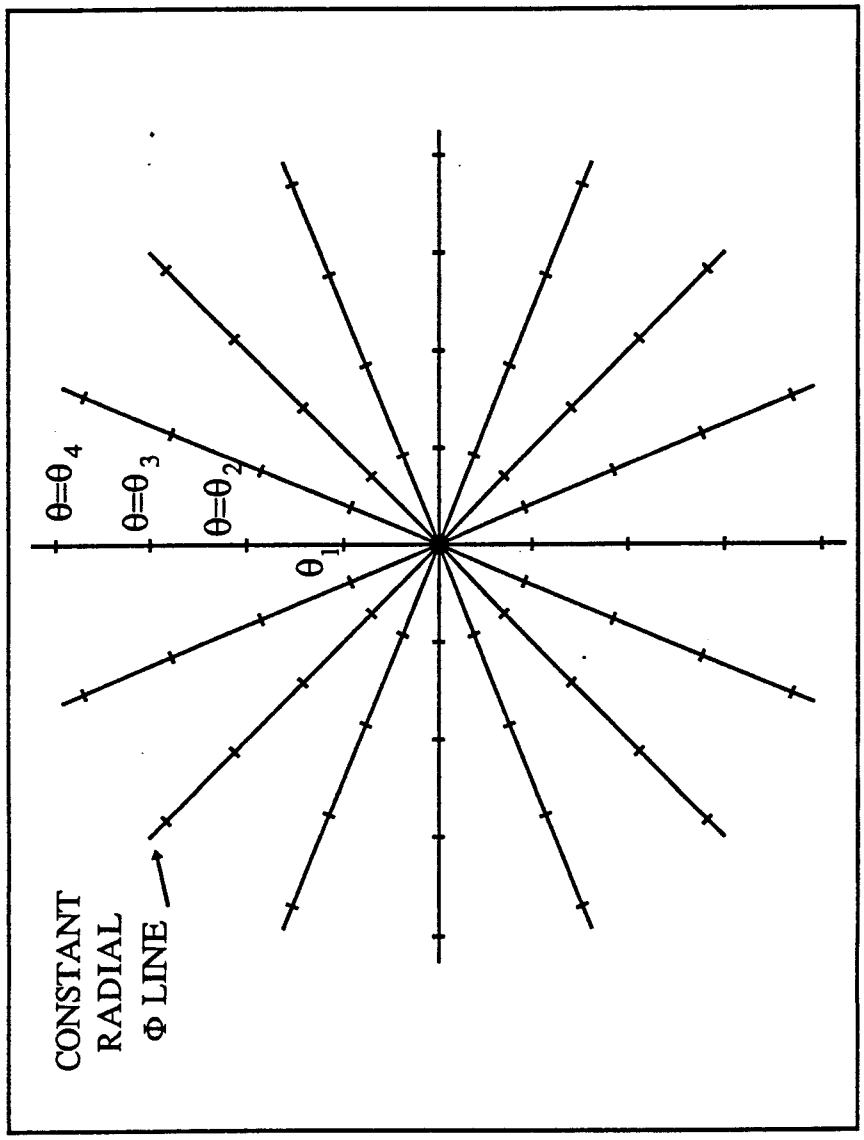


Figure 4: Image Plane

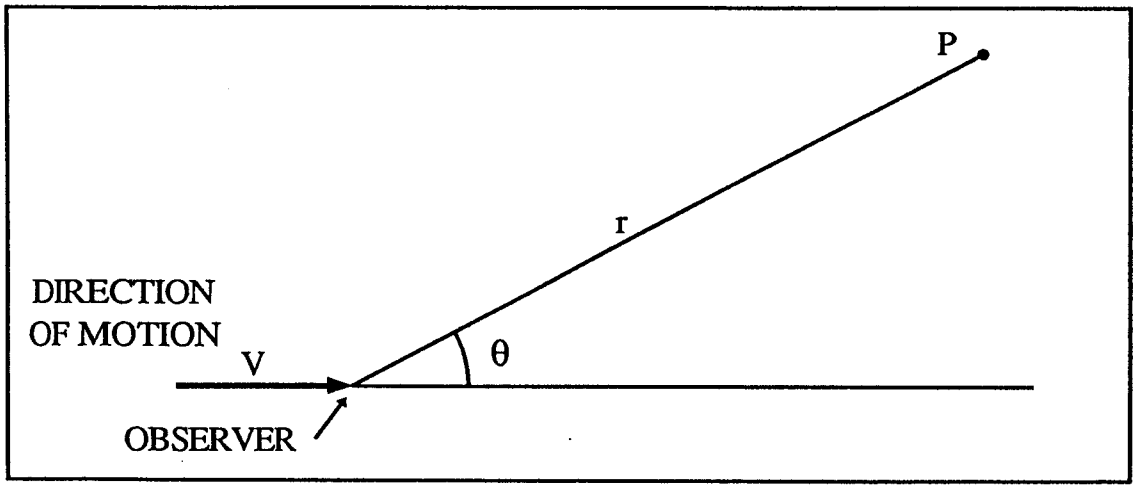


Figure 5: Geometry for Constant Radial Line

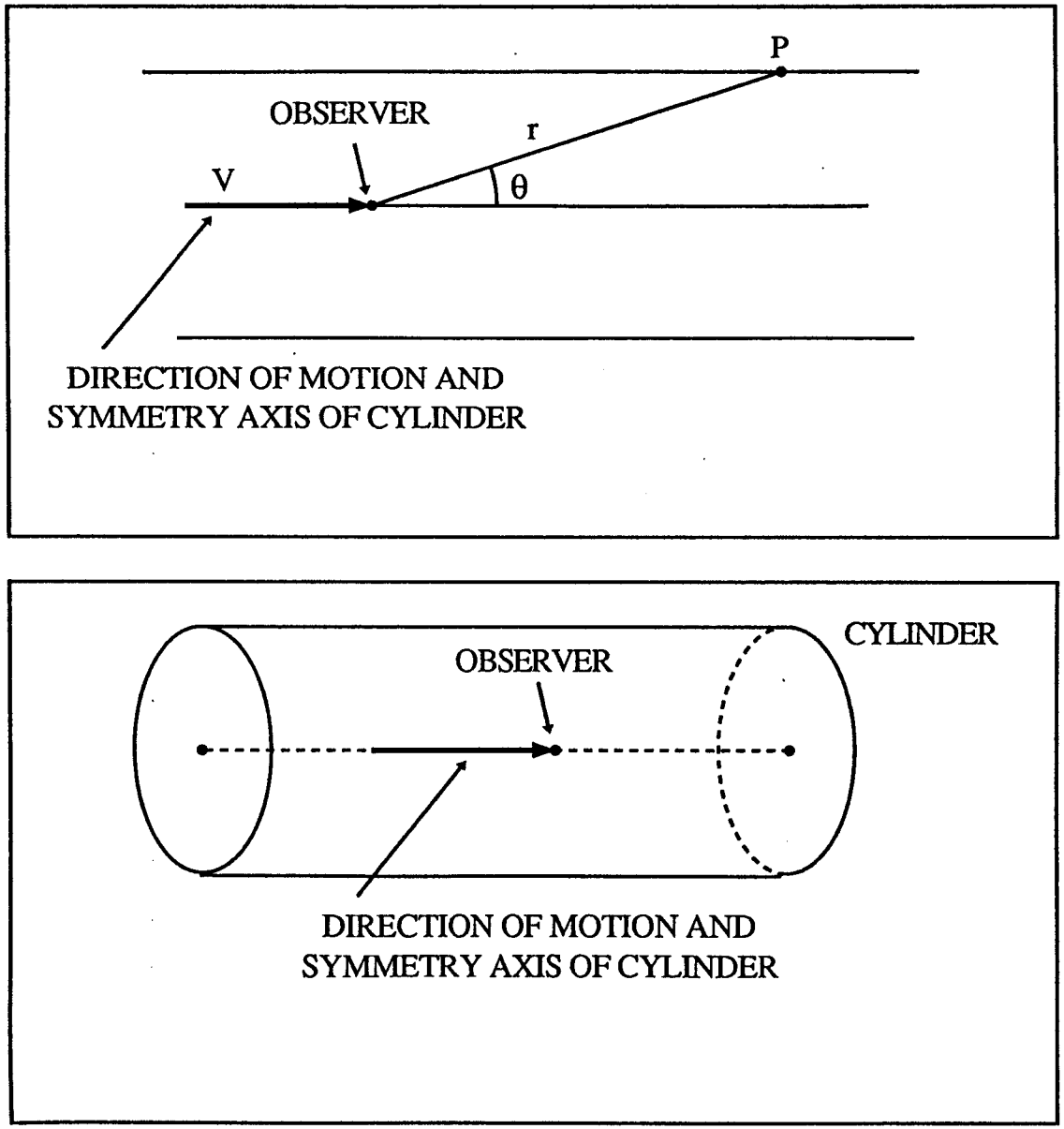
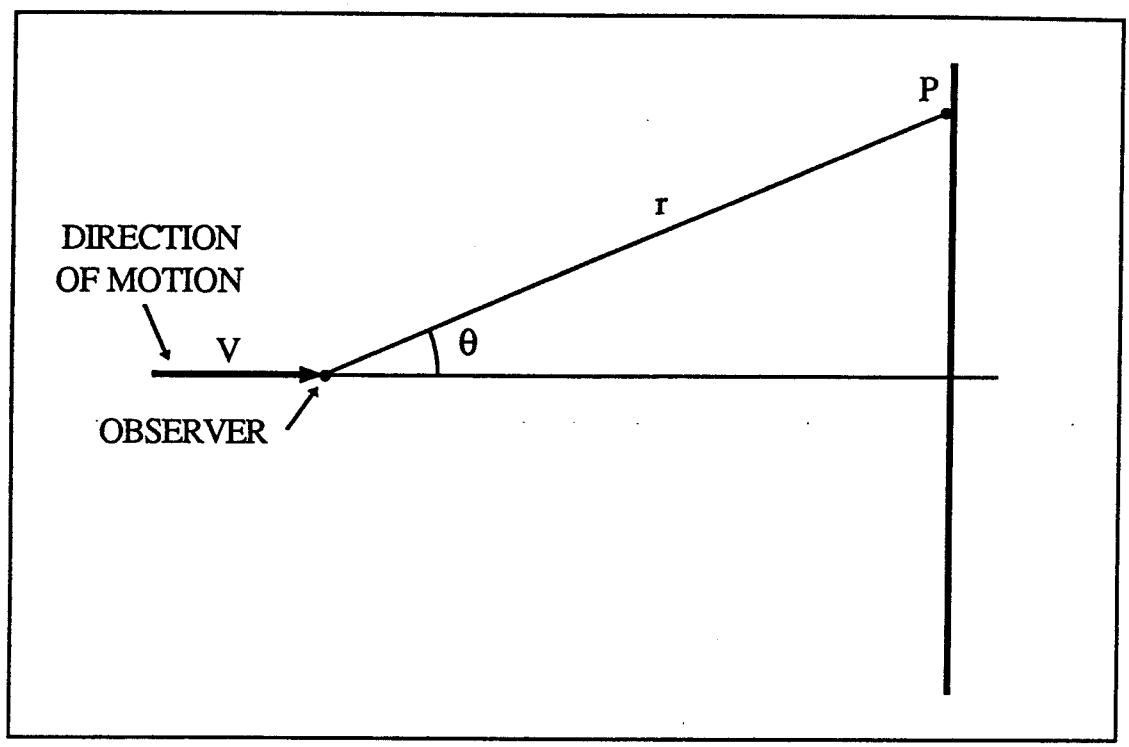


Figure 6: Meaning of the Clearance Invariant

(a) in 2-D

(b) in 3-D



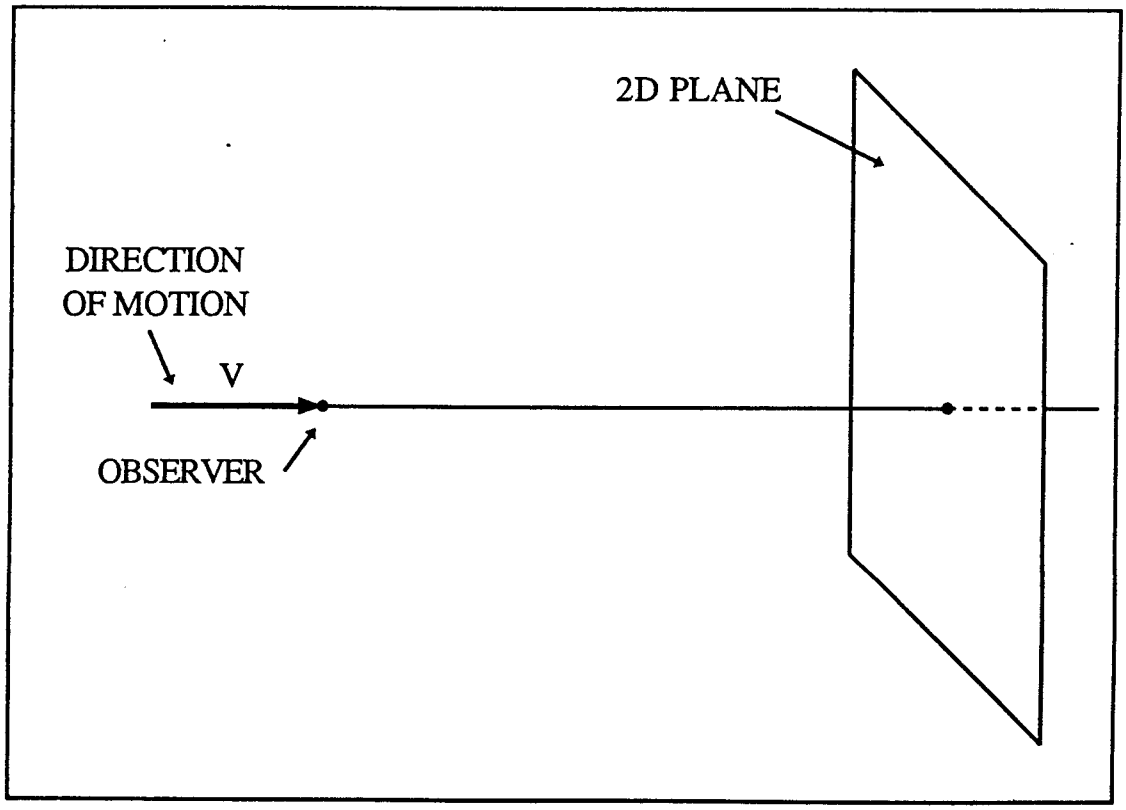


Figure 7: Meaning of the Time-to-Contact Invariant

(a) in 2-D

(b) in 3-D

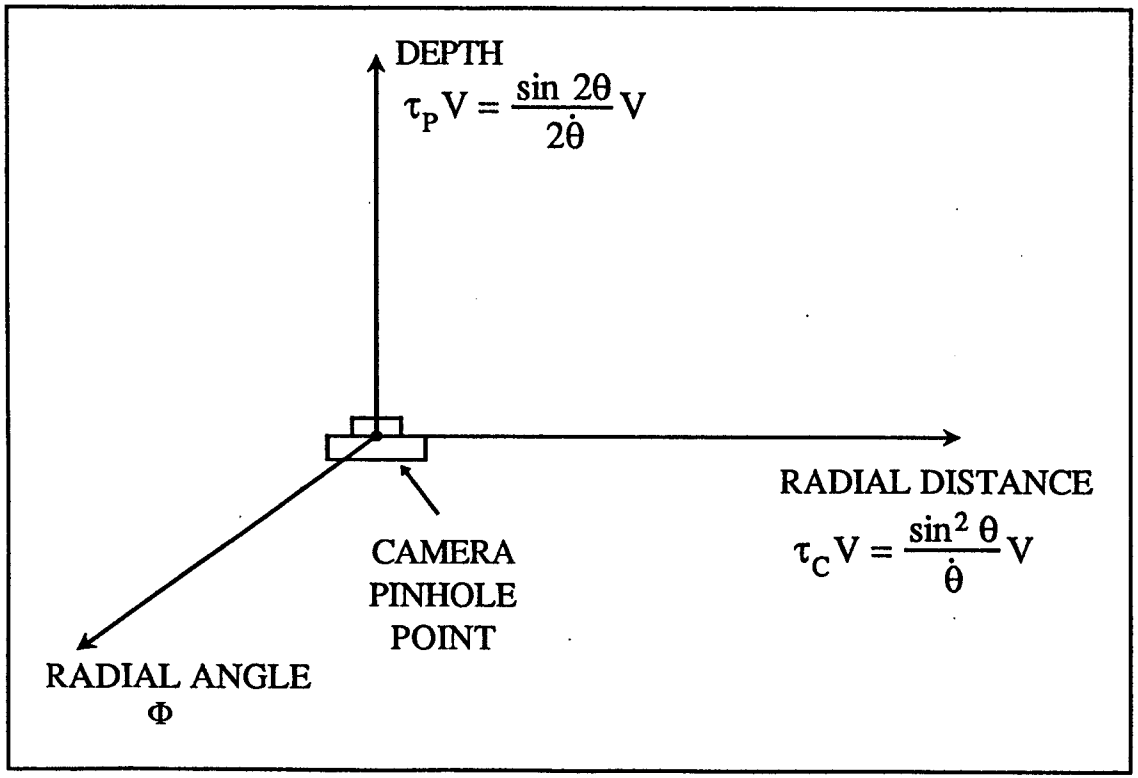
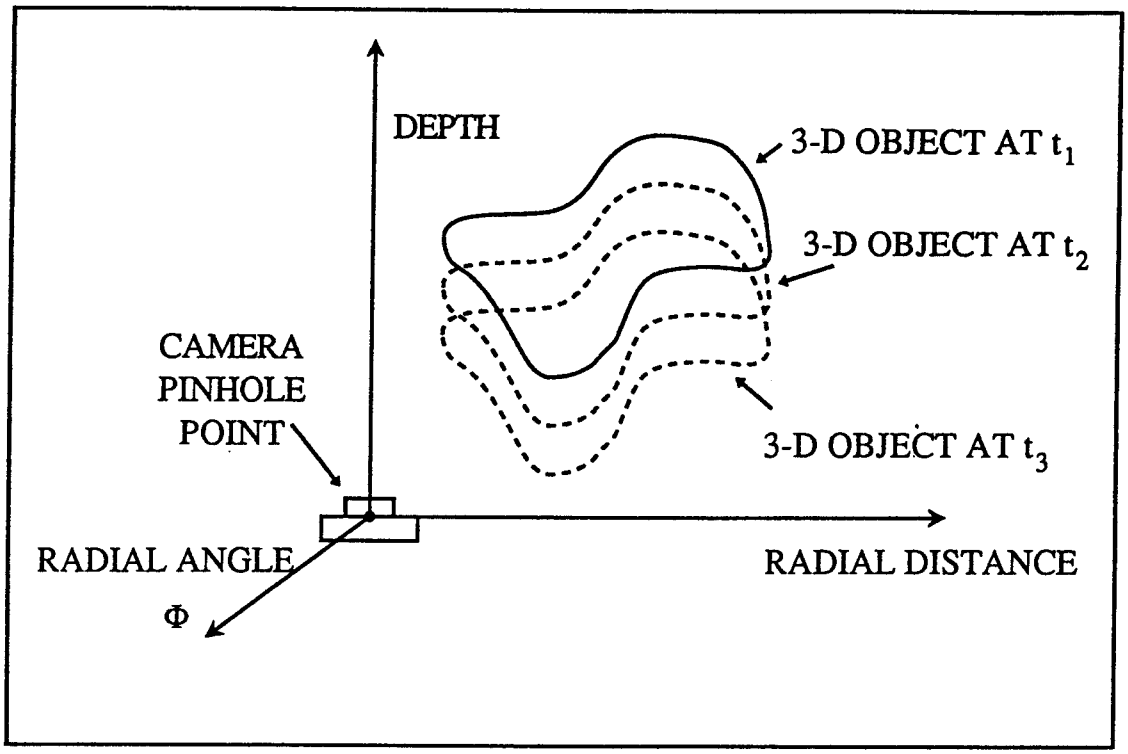


Figure 8: Optical-Flow-Based Coordinate System



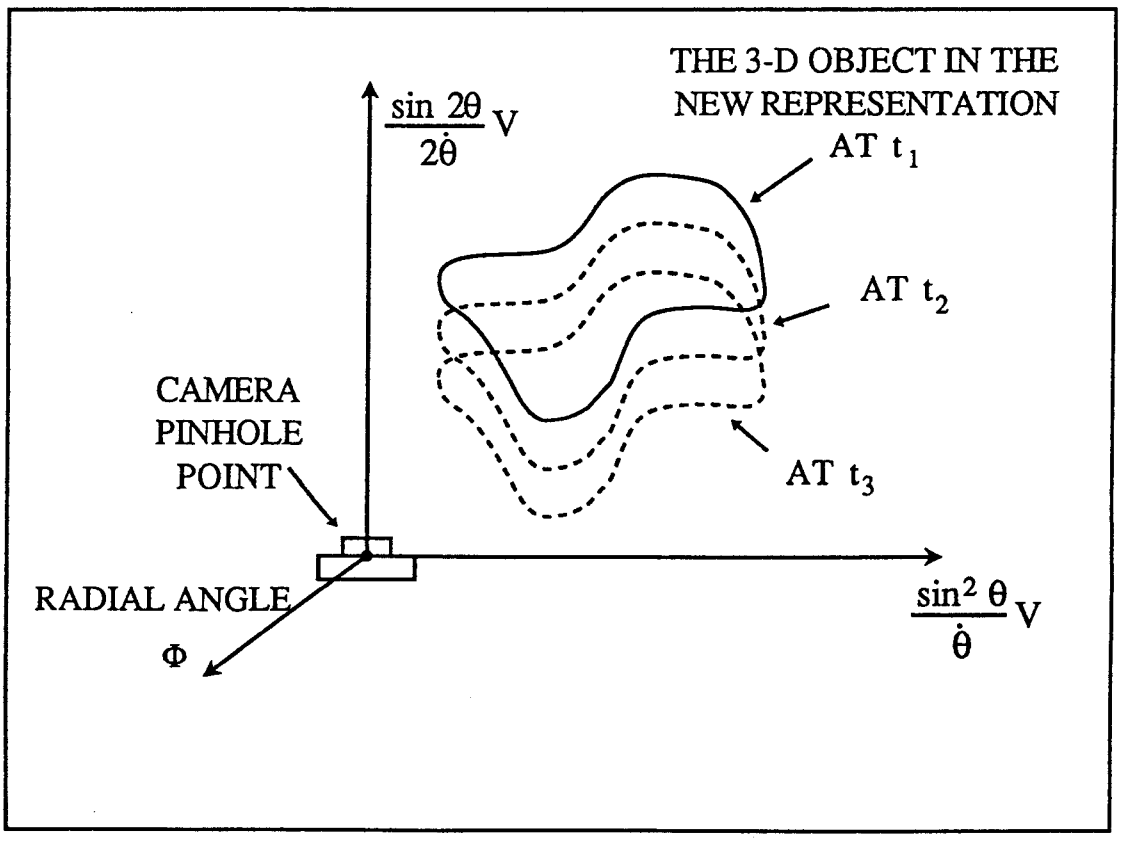


Figure 9: Optical-Flow-Based Coordinate System

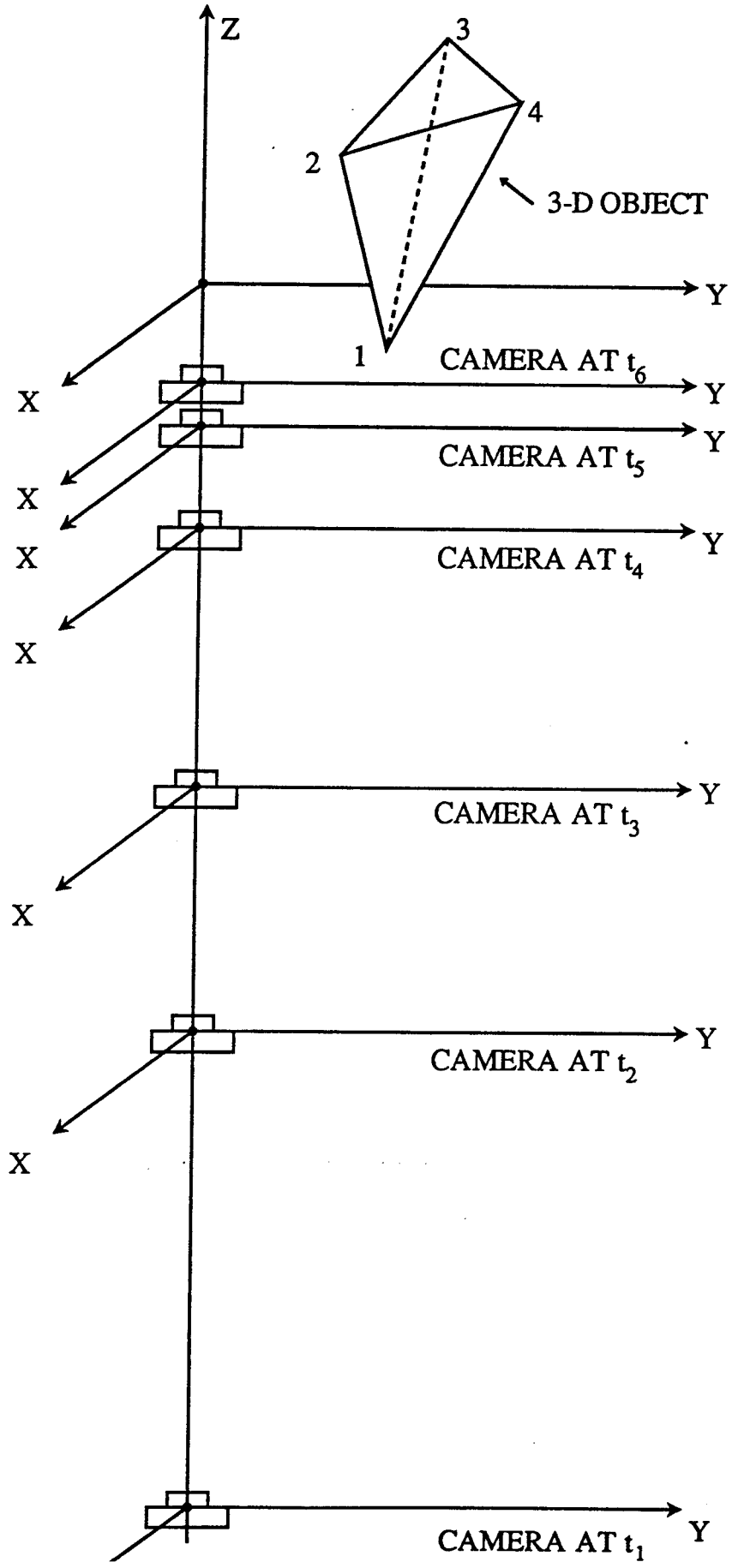
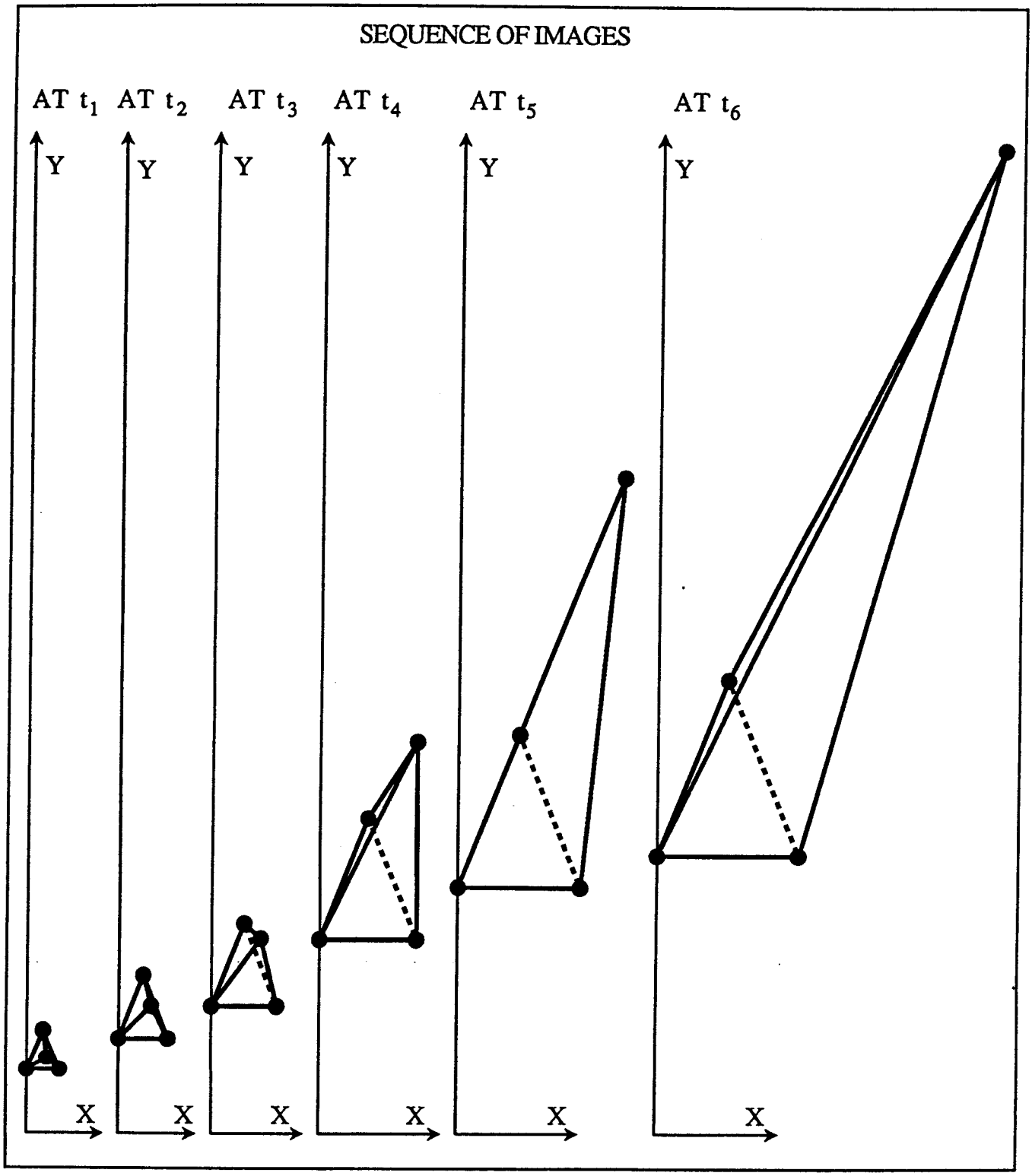
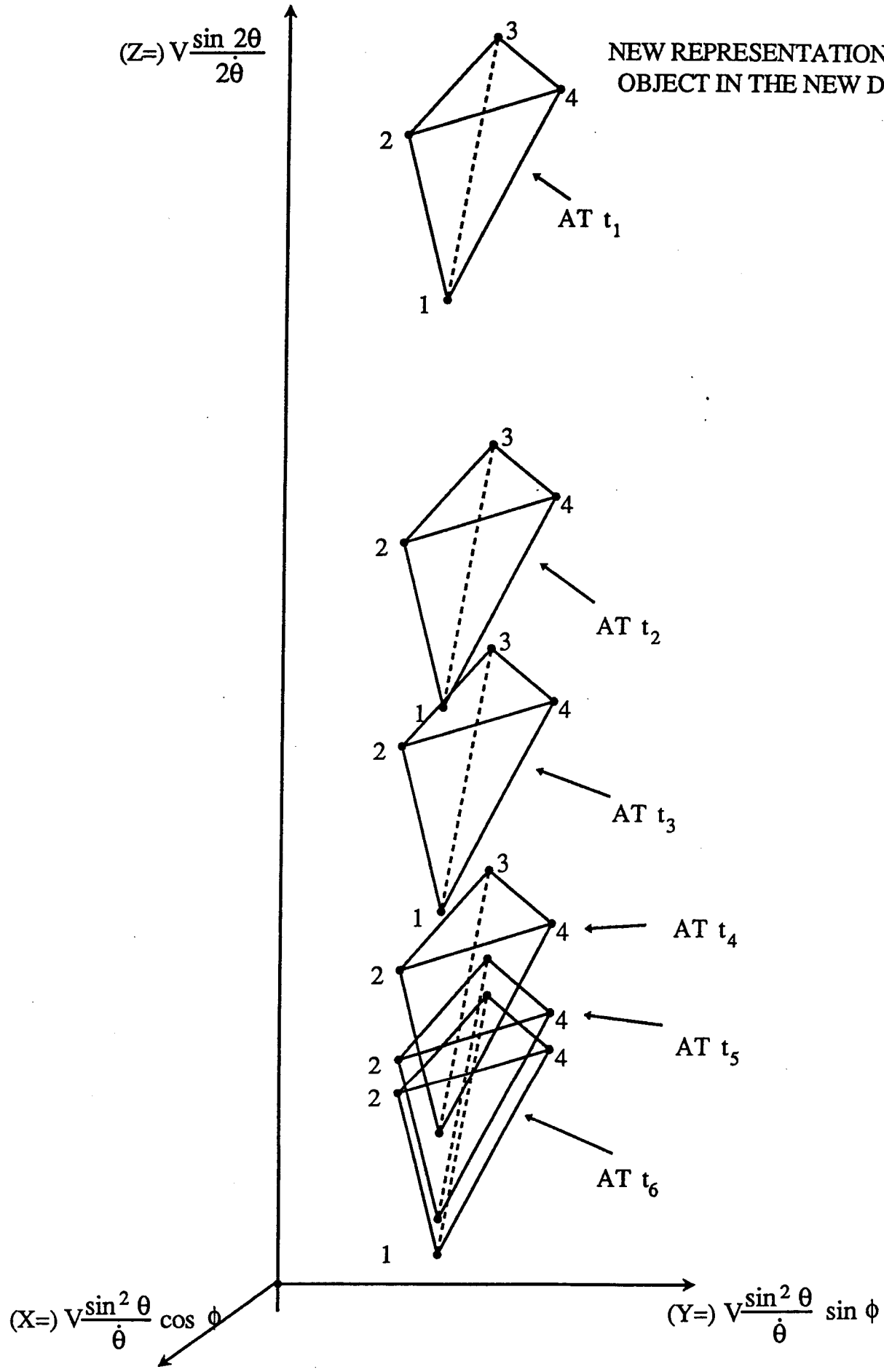


Figure 10: (a) Translating Camera Relative to a 3-D Object
(b) Related Images
(c) The 3-D Object in The New Representation


SEQUENCE OF IMAGES
AT t_1
AT t_2
AT t_3
AT t_4
AT t_5
AT t_6
Y
X


(Z=) $V\frac{\sin 2\theta}{2\dot{\theta}}$
NEW REPRESENTATION OF THE OBJECT IN THE NEW DOMAIN
1
2
3
4
AT t_1
AT t_2
AT t_3
AT t_4
AT t_5
AT t_6
(X=) $V\frac{\sin^2\theta}{\dot{\theta}}\cos\phi$
(Y=) $V\frac{\sin^2\theta}{\dot{\theta}}\sin\phi$

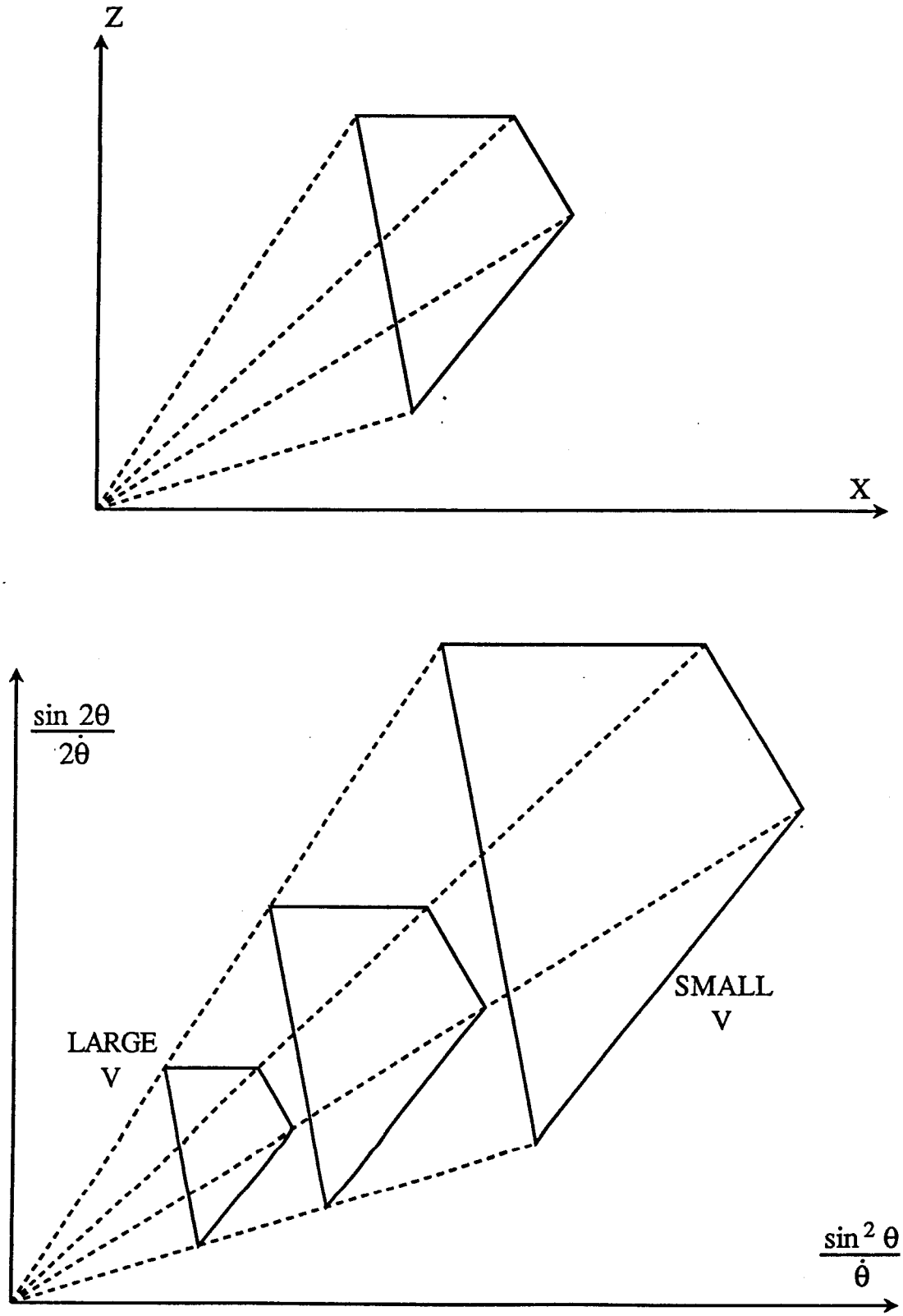


Figure 11: Scaling by V

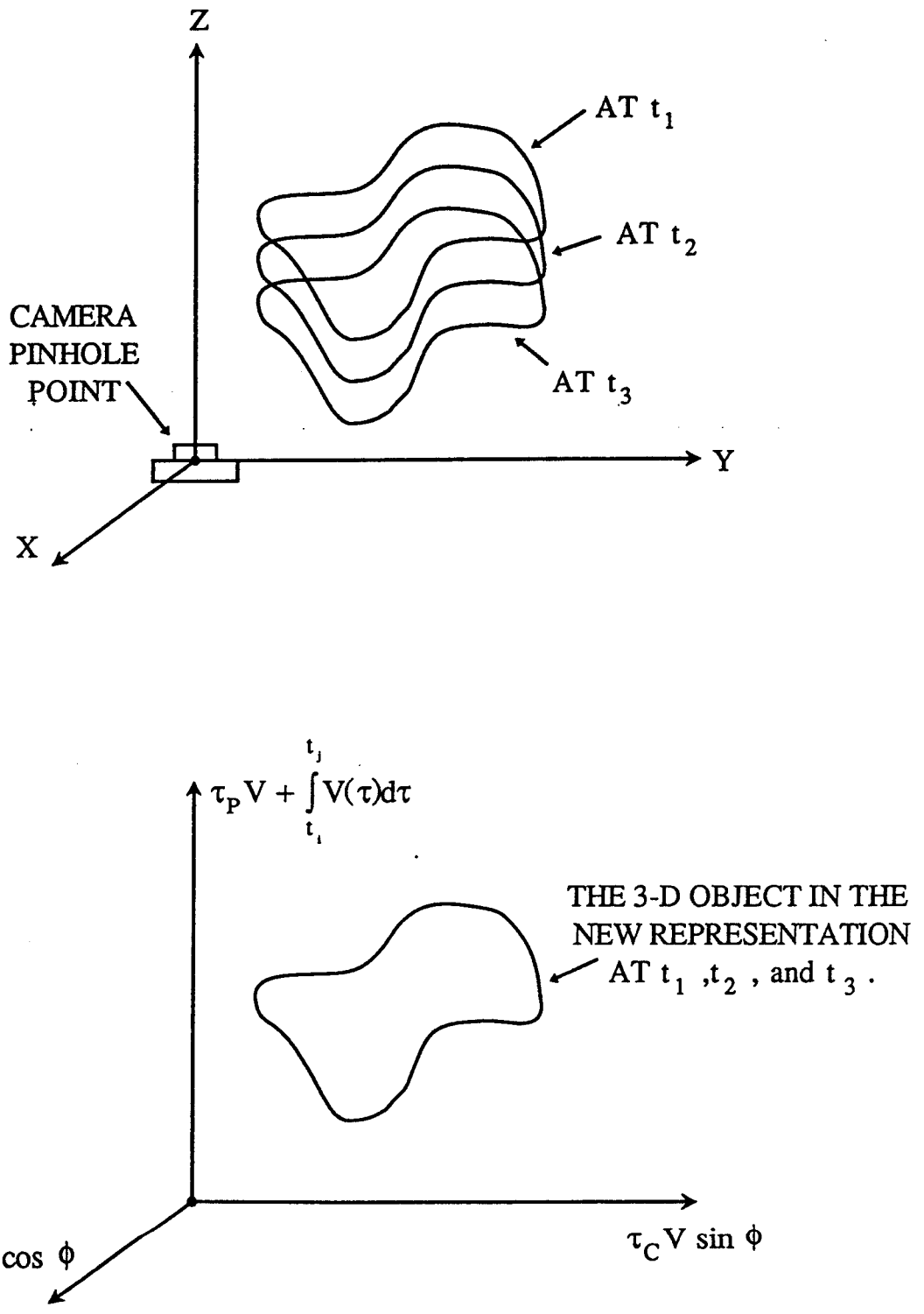


Figure 12: (a) Object Relative to a Camera at Three Different Time Instants

(b) Representation of the Object in The "Frozen" Domain

The problem of constant perception despite varying visual sensations i.e., how despite different images of the same scene, the world is perceived as stationary, has puzzled many researchers in the last five decades. In an attempt to answer this question another basic question arises: Is there an image-sequence transformation under which permanent properties of the environment are preserved during the motion of the eye?

This paper deals with motion-based image transformations that lead to new representations of 3-D objects for a rectilinear motion of an observer. In the new representations a stationary environment is kept invariant or "frozen" in spite of the fact that the images on the retina are continuously changing. Since the 3-D stationary environment is also represented as stationary environment, moving objects can be easily detected and a good estimation of the scene from a long sequence of noisy images can be obtained.

Three basic observations which are related to translational motion of an observer led to the new representations. One is that the radial distance of a point in 3-D space from the camera translational path is kept constant at any instant of time. The second observation is that the relative depth along the translational path between two points is the same at any point in time. The third observation is that points in the image plane move away from the Focus of Expansion (FOE) and towards the Focus of Contraction (FOC).

The new representations are simple. All measurements for these representations are taking place in camera coordinates, only one camera is necessary, and the magnitude of the camera velocity vector needs not be known. They are pixel based i.e., involve local and parallel computations. The mathematics involved is simple, and thus it may be suitable for real time applications.